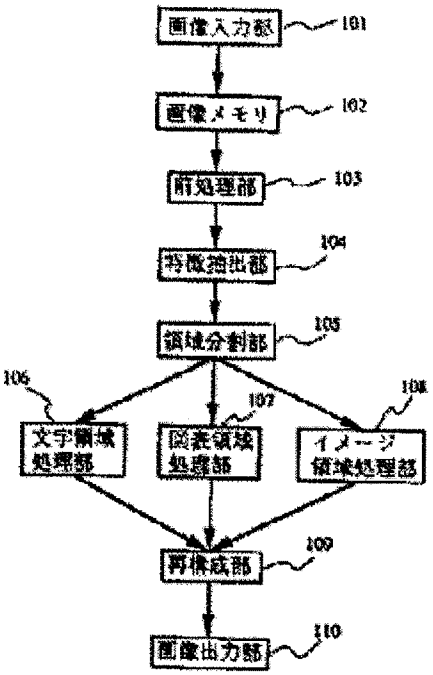
DOCUMENT PICTURE PROCESSOR

Publication number: JP4105178
Publication date: 1992-04-07
Inventor: KUWATA NAOKI
Applicant: SEIKO EPSON CORP
Classification:
- international: ***G06T11/60;* G06T11/60**; (IPC1-7): G06F15/62
- European:
Application number: JP19900223813 19900824
Priority number(s): JP19900223813 19900824

Abstract of **JP4105178**

PURPOSE:To reduce a storage capacity by separate-extracting a character area, a graphic area and an image data from document information fetched in the form of image information, converting the information included in the respective areas into code data suited to the information, after that, reconstituting and outputting them. CONSTITUTION:A picture input part 101, a picture memory 102, a preprocessing part 103, a feature extracting part 104, an area dividing part 105, a character area processing part 106, a graphic area processing part 107, an image area processing part 108, a reconstituting part 109 and a picture output part 110 are provided. Then, an inputted document picture is separated into a character area, a graphic area and an image area, moreover, a character is recognized in the character area, the character and a geometric graphic are recognized in the graphic area, they are converted into code data suited to the respective information and a document is stored. Thus, in the case of storing, a small storage capacity is satisfied.



⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平4−105178

⑤1Int. Cl.[5]	識別記号	庁内整理番号	⑬公開　平成4年(1992)4月7日
G 06 F 15/62	325 P	8125−5L	

審査請求　未請求　請求項の数 1　(全6頁)

⑤4発明の名称　文書画像処理装置

⑳特　願　平2−223813
㉒出　願　平2(1990)8月24日

⑦2発明者　鍬田　直樹　長野県諏訪市大和3丁目3番5号　セイコーエプソン株式会社内
⑦1出願人　セイコーエプソン株式会社　東京都新宿区西新宿2丁目4番1号
⑦4代理人　弁理士　鈴木　喜三郎　外1名

明細書

1．発明の名称

文書画像処理装置

2．特許請求の範囲

文字・図表・イメージ領域を少なくとも一つ以上含む文書をイメージ情報として取り込む画像入力部と、前記画像入力部により取り込まれたイメージ情報から雑音を除去し2値化する前処理部と、前記イメージ情報における前記文字・図表・イメージ領域の持つ特徴を抽出する特徴抽出部と、前記特徴抽出部で抽出された特徴に基づき前記イメージ情報を文字領域・図表領域・イメージ領域に分割する領域分割部と、前記文字領域部の文字を認識する文字領域処理部と、前記図表領域内の幾何図形および文字を認識する図表領域処理部と、前記イメージ領域内のイメージ情報を加工するイメージ領域処理部と、前記3処理部からのデータを再構成する再構成部と、再構成された文書情報を出力する画像出力部とを具備したことを特徴とする文書画像処理装置。

3．発明の詳細な説明

[産業上の利用分野]

本発明は、イメージ情報の形式で取り込まれた文書情報から文字領域と図表領域とイメージ領域を分離抽出し、各領域に含まれる情報をその情報に適するコードデータに変換後、再構成して出力する文書画像処理装置に関する。

[従来の技術]

一般的に、紙に書かれた文書情報を保存する場合、イメージスキャナ等で取り込んだ画像をそのままイメージ情報として扱い、外部記憶装置等に保存している。また、文字領域のみを抽出した後文字領域について文字認識を行い文字コード化して保存する場合は、入力文書画像に対して使用者が、文字領域を人手により指定する必要があった。

[発明が解決しようとする課題]

以上述べたように、イメージ情報の形で保存するタイプでは、記憶容量が膨大になり、また、文

書の一部を書き直したりする編集作業が行えなかった。そして、領域指定するタイプでは、人間が常にその場に居て、指示する必要があり、手間がかかっていた。

そこで本発明は、上記の問題点を解決するためのもので、人手に頼らず入力された文書画像を文字領域・図表領域・イメージ領域に分離し、さらに文字領域においては文字を認識し、図表領域においては文字および幾何図形を認識し、それぞれの情報に適したコードデータに変換して、文書を保存する装置を提供することを目的とする。

［課題を解決するための手段］

本発明の文書画像処理装置は、文字・図表・イメージ領域をを少なくとも一つ以上含む文書をイメージ情報として取り込む画像入力部と、この画像入力部により取り込まれたイメージ情報から雑音を除去し２値化する前処理部と、イメージ情報における文字・図表・イメージ領域の持つ特徴を抽出する特徴抽出部と、この特徴抽出部で抽出された特徴に基づき前述のイメージ情報を文字領域・図表領域・イメージ領域に分割する領域分割部と、文字領域部の文字を認識する文字領域処理部と、図表領域内の幾何図形および文字を認識する図表領域処理部と、イメージ領域内のイメージ情報を加工するイメージ領域処理部と、これらの各処理部からのデータを再構成する再構成部と、再構成された文書情報を出力する画像出力部とを具備したことを特徴とする。

［実施例］

以下本発明について図面に基づいて説明する。第１図は本発明の文書画像処理装置の構成を示すブロック図である。１０１は文書画像をイメージ情報として取り込む画像入力部であり、スキャナもしくはカメラ等を用いる。あらかじめ画像が収納されている光ディスク等を使用する場合は、これに対応する再生装置になる。１０２は取り込んだイメージ情報を一時的に保存する画像メモリである。１０３はイメージ情報に含まれる雑音の除去、２値化を行う前処理部である。雑音の除去には、メジアンフィルタ等を用いて孤立雑音を除去する。もし、入力画像が傾いていたときは後の処理をやりやすくするために、この部分で傾斜角の補正を行う。１０４は文字領域・図表領域・イメージ領域を分離するための特徴量を抽出する特徴抽出部である。特徴の抽出法については後述する。１０５は、特徴抽出部１０４で抽出された特徴量に基づいて入力されたイメージ情報を文字領域・図表領域・イメージ領域に分割する領域分割部である。１０６は分割された文字領域内で、文字列の抽出、一文字の切り出し、切り出した文字の認識を行う文字領域処理部である。１０７は図表領域内の幾何図形および文字を抽出した後、認識を行う図表領域処理部である。１０８はイメージ領域と判定された部分をイメージデータのまま、もしくは圧縮処理をするイメージ領域処理部である。１０９は、コードデータ化された領域とイメージデータの領域を紙面上に再構成する再構成部である。１１０は再構成された文書情報を出力する画像出力部で、具体的には印画装置・表示装置・外部記憶装置がこれに該当する。

次に、入力されたイメージ情報の特徴抽出の方法について説明する。まず、入力画像を縦ｍ個、横ｎ個の画素ごとにグループ化する。そして、各グループ（ｍ×ｎ画素）中に存在する黒点の数を計数する。この操作を入力画像全体に対して行う。第２図は入力されたイメージ情報をｍ×ｎ画素のグループに分割し、そのなかに存在する黒画素の数を入力画像全面に渡って計数し、密度（黒画素数）を横軸に、その出現頻度（度数）を縦軸にとったヒストグラムを示す図である。一般的に、図表領域は白い部分が多く密度は低くなる。一方、イメージ領域は黒い部分が多く密度は高くなる。文字領域はこの中間に位置する。図に示されたように、適当なしきい値（ｔ１、ｔ２、ｔ３）で分離された領域を密度の低い順番に０、１、２、３と番号を振ると、０は何も書かれていない空白領域、１は図表領域、２は文字領域、３はイメージ領域というように分割することができる。この例の場合は、文字領域の面積が大きい文書を標本として用いたので、文字領域に対応する部分の度数

が多くなっている。

第３図は、第２図に示した方法により入力文書に対して、ｍ×ｎ画素ごとにラベル付けを行った一例を示す図である。１とラベルがつけられた領域が図表領域、２が文字領域、３がイメージ領域に対応する。このようにして、同じラベルの付いた領域をグループ化することにより、領域分割を行う。

第４図は、文字領域処理部１０６の詳細を示すブロック図である。４１は文字領域内の文字列を抽出する文字列抽出部、４２は抽出された文字列から一文字を切り出す文字抽出部、４３は抽出された文字を文字認識用辞書４４を参照して、認識を行う文字認識部である。ここで認識が行えなかった文字に関しては、イメージデータのまま次の単語照合部４５へ送られる。単語照合部では、認識された文字が、単語として意味をもつかどうか単語辞書４６を参照して、もし文字の誤認識により意味のない単語が存在した場合は、訂正の可能なものについては正しい単語に変換する。文字認識部で認識できなかった文字についても、単語辞書を参照することにより確定できるものについては、この部分で決定する。単語照合部で確定できなかった文字については、イメージのまま残しておく。４７は、認識された文字について、これをコード化する文字コード化部である。４８は、上記の部分でコードデータ化された文字列を紙面上で再構成するために必要な情報を付加する頁書式付加部である。例えば、第５図に示されるように、紙面の左上を原点として、縦方向にＸ軸を、横方向にＹ軸をとったとき、（ｘ１、ｙ１）と（ｘ２、ｙ２）で囲まれた領域に文字列が存在するとする。このとき、この領域を示す頁書式は、例えば第６図（ａ）に示したようになる。この例では、"Ｃｈａｒ"が、この領域が文字領域であることを、文字領域の存在位置が（ｘ１、ｙ１）と（ｘ２、ｙ２）で囲まれた矩形内であること、文字の種類が明朝体であること、文字の大きさが１０ポイントであり、認識した文字列が|@@……@@|で示される内容であることをそれぞれ表している。

この情報を基に、文字領域内の情報が印字装置や表示装置に出力されたり、あるいは外部記憶装置に保存される。

第７図は、図表領域処理部１０７の詳細を示すブロック図である。７１は図表領域内の幾何図形を抽出する幾何図形抽出部で、抽出法には、Ｈｏｕｇｈ変換・黒画素の連結成分抽出等を用いる。７２は図表領域に含まれる文字を抽出する文字抽出部、７３は抽出された文字を文字認識用辞書７４を参照して、認識を行う文字認識部である。ここで認識が行えなかった文字に関しては、イメージデータのまま次の単語照合部７５へ送られる。単語照合部では、認識された文字が、単語として意味をもつかどうか単語辞書７６を参照して、もし文字の誤認識により意味のない単語が存在した場合は、訂正の可能なものについては正しい単語に変換する。文字認識部で認識できなかった文字についても、単語辞書を参照することにより確定できるものについては、この部分で決定する。単語照合部で確定できなかった文字については、イメージのまま残しておく。７７は、認識された文字について、これをコード化する文字コード化部である。７８は、上記の部分でコードデータ化された文字列を紙面上で再構成するために必要な情報を付加する頁書式付加部である。例えば、第５図に示されるように、紙面の左上を原点として、縦方向にＸ軸を、横方向にＹ軸をとったとき、（ｘ３、ｙ３）と（ｘ４、ｙ４）を結ぶ直線が存在するとする。このとき、この直線を示す頁書式は、例えば第６図（ｂ）に示したようになる。この例では、"Ｌｉｎｅ"が、この図形が直線であることを、直線の存在位置が（ｘ３、ｙ３）と（ｘ４、ｙ４）を結ぶ領域であること、線の種類が実線であること、線の幅が０．５ｍｍであることをそれぞれ表している。第７図において、７２から７８で示される部分については、文字領域処理部１０６に含まれるものと共用してもよい。

本発明の応用例としては、以下のものが考えられる。電子ファイリングシステムにおいて、入力画像を領域分割し、コード化することによって、

データの圧縮ができ、記憶容量が縮小できる。デスクトップパブリッシングと組み合わせることにより、入力画像の文章や図形を書き換えて別の文書を作成するのに利用することができる。機械翻訳を行う際、従来キーボードなどを使用して人が入力していた文書入力の自動化を図ることができる。複写機において、従来イメージ情報のまま複製を繰り返すとと、雑音等の影響で文字や図形が不鮮明になりついには読み取れなくなったが、一度この文書処理装置を通し、コード化できる部分についてコード化することにより、コード化された部分については、何回複写を繰り返しても常に鮮明な画像を得ることができる。また、同様の理由でファクシミリの入力画像の処理に利用すると、画像が鮮明になり、かつ伝送容量の圧縮につながる。

［発明の効果］

以上述べたように、本発明の文書画像処理装置を用いると、従来イメージデータとして取り扱っていた文字および図形を認識することにより、これに適したコードデータに直すので、保存する場合、記憶容量が少なくて済む。またコードデータに変換されているので、一部分の文字・図形等を変更したり、再利用したりする編集作業が行える。さらに、文字・図表・イメージ領域を自動的に分離抽出しているので処理の省力化が可能になるだけでなく、あらかじめプログラムを設定しておくことにより、欄外に存在するロゴマークを消すとか、文章だけ、図形だけの保存といったトリックプレイも行える。

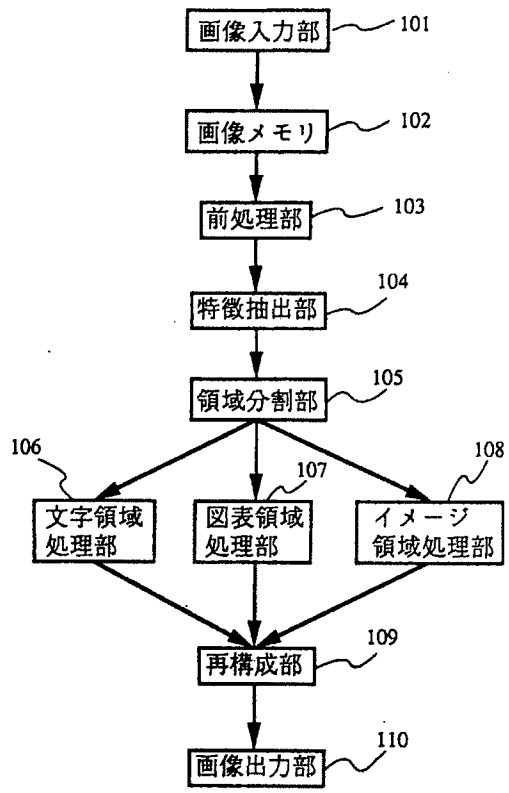
4．図面の簡単な説明

第1図は本発明の文書画像処理装置の構成を示すブロック図、第2図は各領域の黒画素の分布を示す図、第3図は入力画像を領域分割したときの図、第4図は本発明の文字領域処理部のブロック図、第5図は入力画像の一例を示す図、第6図は第5図を頁書式で表現した図、第7図は本発明の図表領域処理部のブロック図である。

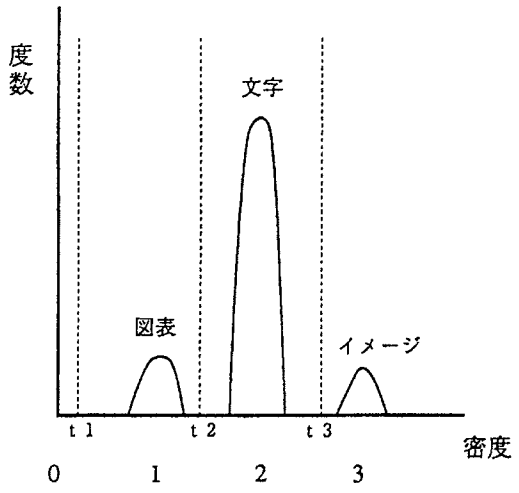
１０１…画像入力部、１０２…画像メモリ、１０３…前処理部、１０４…特徴抽出部、１０５…領域分割部、１０６…文字領域処理部、１０７…図表領域処理部、１０８…イメージ領域処理部、１０９…再構成部、１１０…画像出力部、４１…文字列抽出部、４２・７２…文字抽出部、４３・７３…文字認識部、４４・７４…文字認識用辞書、４５・７５…単語照合部、４６・７６…単語辞書、４７・７７…文字コード化部、４８・７８…頁書式付加部、７１…幾何図形抽出部

出願人　セイコーエプソン株式会社

代理人　弁理士鈴木喜三郎（他１名）



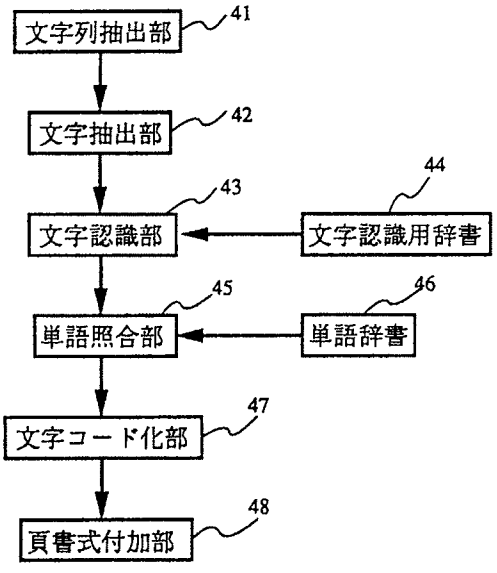
第1図



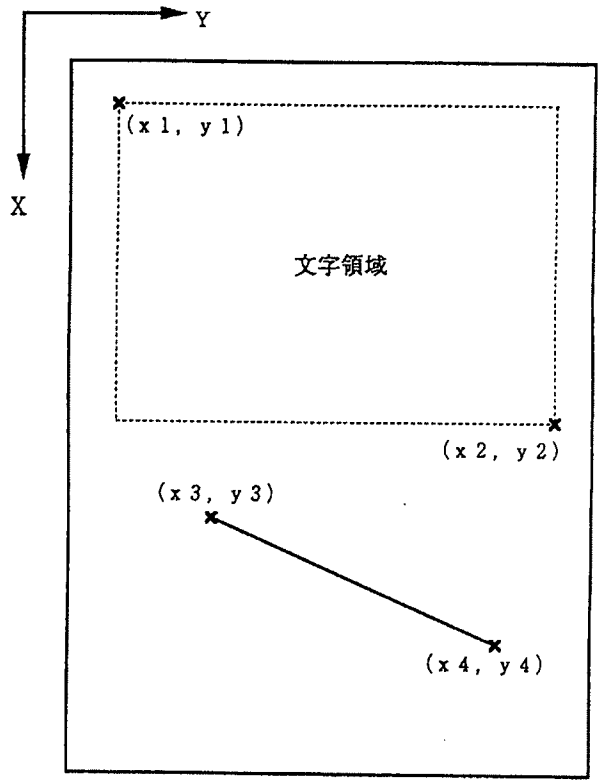
第2図

0	0	0	0	0	0	0	0	0	0	0
0	2	2	2	2	2	2	2	2	2	0
0	2	2	2	2	2	2	2	2	2	0
0	2	2	2	2	2	2	2	2	2	0
0	2	2	2	2	2	2	2	2	2	0
0	2	2	2	2	2	2	2	2	2	0
0	2	2	2	2	2	2	2	2	2	0
0	0	0	0	0	0	0	0	0	0	0
0	3	3	3	3	0	1	1	1	1	0
0	3	3	3	3	0	1	1	1	1	0
0	3	3	3	3	0	1	1	1	1	0
0	3	3	3	3	0	1	1	1	1	0
0	3	3	3	3	0	1	1	1	1	0
0	3	3	3	3	0	0	0	0	0	0
0	3	3	3	3	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

第3図



第4図



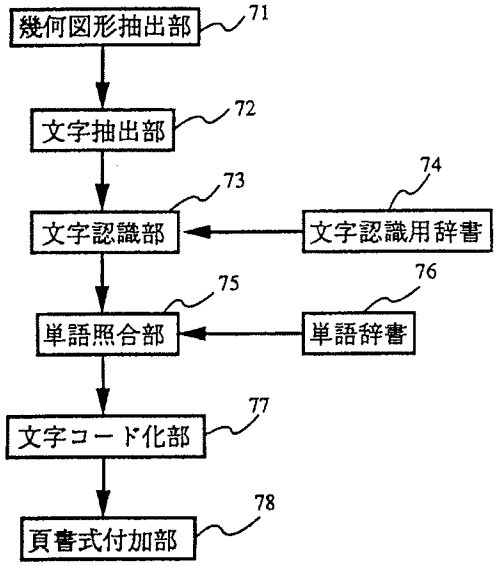
第5図

```
Begin:Char [
位置 |(x1, y1) − (x2, y2)|
文字種 |明朝|
文字サイズ |10ポイント|
文字列 |@@……@@|
] :End
```

第6図（a）

```
Begin:Line [
位置 |(x3, y3) − (x4, y4)|
線種 |実線|
線幅 |0. 5|
] :End
```

第6図（b）



第7図